Introduction

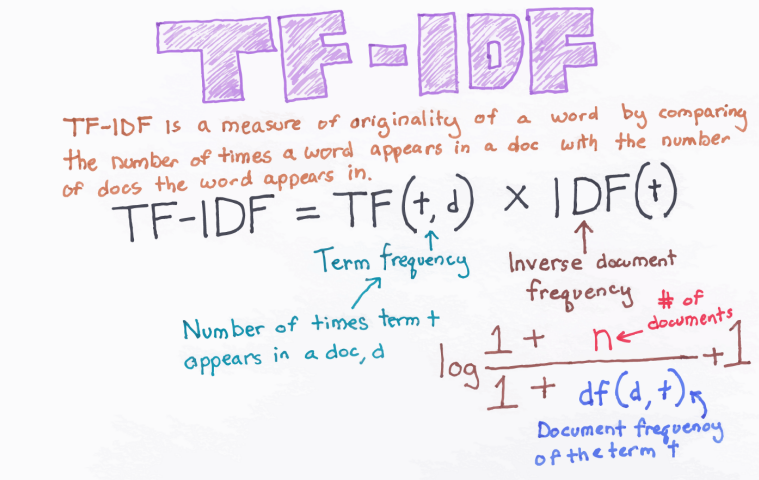
Log Analysis is the process of making sense of computer generated log message also known as log events or simple logs . Log analysis provide us clear picture of what has happened across the infrastructure

Method Followed:

- First step is to connect the elastic search to the python script, For this we used elastic search library to connect and extract data for Analysis
  Reference : Link1

- Next step in our analysis is text cleaning. Here by looking at the logs we removed some strings which are not required in our analysis like removing date and time ,removing Ip addresses , removing any special character from the log text etc.
- Tokenization :
  In this step we are breaking the raw text into small chunks .Tokenization breaks the raw text into words , sentences called tokens . These tokens help in understanding the context or developing the model for NLP
  Reference : Link2
- Stemming :
  It is a process of producing morphological variants of root/base word. Stemming is an important part of the pipelining process in Natural language processing. The input to the stemmer is tokenized words
  Reference : Link3

- TFIDF vectorization:
  This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.



  Reference :Link3

- In the next step using the vector array from the TF IDF vectorization we will apply the DBSCAN Algorithm on it. The algorithm uses density to cluster the data points. We can also identify the noise very well using this algorithm
  Reference : Link4
  - We also applied K means algorithm on the matrix obtained from the vectorization step
    In K means we group the unlabeled dataset into the different clusters .
    It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties.
    Reference : Link 5
- Used MDS (Multidimensional Scaling) to reduce the dimensionality which will help us to visualize the data in 2D plane
  Reference : Link 6
- Anomaly Detection with Histogram-based Outlier Score (HBOS) is used for calculating score
  In HBOS we can define the univariate outlier score based on the histogram of a variable. We can also add up all the the univariate outlier scores to get the multivariate outlier score for a data point
  Reference : Link 7
- Isolation Forest : The existing approaches to anomaly detection find the norm first, then identify observations that do not conform to the norm. They propose the Isolation Forest as an alternative approach — explicitly isolating anomalies instead of profiling normal data points. Anomalies are isolated closer to the root of the tree
  Reference : Link 8